

HYBRID MACHINE LEARNING FRAMEWORK FOR FORECASTING AND EVALUATING UNIVERSITY RANKINGS

Nurzhigit Smailov^{1,3}, Nursultan Kuldeyev², Akezhhan Sabibolda^{1,3}, Raigul Ustemirova⁴

¹Satbayev University, Department of Radio Engineering, Electronics and Space Technologies, Almaty, Kazakhstan, ²Satbayev University, Department of Information Systems, Almaty, Kazakhstan, ³Institute of Mechanics and Machine Science Named by Academician U.A. Dzholdasbekov, Almaty, Kazakhstan, ⁴International University of Transportation and Humanities, Department of Traffic Management, Transport Management and Logistics, Almaty, Kazakhstan

Abstract. The growing importance of university rankings in higher education requires evaluation methods that are both predictive and interpretable. Traditional ranking approaches are usually based on fixed weighted indicators, while purely data-driven machine learning models often provide limited transparency for institutional decision-making. This paper proposes a hybrid analytical-machine learning framework for evaluating and forecasting university rankings using multidimensional performance indicators. The framework combines an interpretable weighted aggregation model, an XGBoost-based residual correction mechanism, an LSTM forecasting module, and SHAP-based explainability analysis. The empirical dataset was constructed as a university-year panel covering 20 Kazakhstani universities over the 2019–2024 period, resulting in 120 observations. The input indicators included publication output, citation impact, h-index, international collaboration, and research funding. The models were evaluated using a temporal validation protocol, with 2019–2022 used for training, 2023 for validation, and 2024 as the independent test subset. The results show that the hybrid model reduced RMSE from 0.185 to 0.092 compared with the analytical model and from 0.124 to 0.092 compared with the standalone XGBoost model. Similar improvements were observed for MAE, MAPE, and ranking error. However, the remaining deviations indicate that the model should be interpreted as reducing, rather than eliminating, prediction uncertainty. SHAP analysis showed that citation impact and publication output had the widest feature-contribution distributions, while all indicators demonstrated both positive and negative effects depending on the observation. The proposed framework can support institutional analytics by combining ranking evaluation, forecasting, and interpretable identification of key performance drivers.

Keywords: university ranking, hybrid machine learning, analytical modelling, forecasting, decision support system, feature importance

HYBRYDOWY MODEL UCZENIA MASZYNOWEGO DO PROGNOZOWANIA I OCENY RANKINGÓW UNIWERSYTETÓW

Streszczenie. Rosnące znaczenie rankingów uczelni wyższych w szkolnictwie wyższym wymaga zastosowania metod oceny, które są zarówno predykcyjne, jak i zrozumiałe. Tradycyjne podejścia do tworzenia rankingów opierają się zazwyczaj na stałych, ważonych wskaźnikach, podczas gdy modele uczenia maszynowego oparte wyłącznie na danych często zapewniają ograniczoną przejrzystość w procesie podejmowania decyzji przez instytucje. W niniejszym artykule zaproponowano hybrydową strukturę łączącą analizę z uczeniem maszynowym, służącą do oceny i prognozowania rankingów uniwersytetów przy użyciu wielowymiarowych wskaźników wydajności. Struktura ta łączy interpretowalny model agregacji ważonej, mechanizm korekcji reszt oparty na XGBoost, moduł prognozowania LSTM oraz analizę wyjaśnialności opartą na SHAP. Zbiór danych empirycznych został skonstruowany jako panel roczno-uczelniany obejmujący 20 kazachstańskich uczelni w latach 2019–2024, co dało 120 obserwacji. Wskaźniki wejściowe obejmowały liczbę publikacji, wpływ cytowań, indeks h, współpracę międzynarodową oraz finansowanie badań naukowych. Modele oceniono przy użyciu protokołu walidacji czasowej, w ramach którego lata 2019–2022 wykorzystano do szkolenia, rok 2023 do walidacji, a rok 2024 jako niezależny podzbiór testowy. Wyniki pokazują, że model hybrydowy zmniejszył wskaźnik RMSE z 0,185 do 0,092 w porównaniu z modelem analitycznym oraz z 0,124 do 0,092 w porównaniu z samodzielnym modelem XGBoost. Podobną poprawę zaobserwowano w przypadku wskaźników MAE, MAPE oraz błędów rankingowego. Pozostałe odchylenia wskazują jednak, że model ten należy interpretować jako zmniejszający, a nie eliminujący niepewność prognoz. Analiza SHAP wykazała, że wpływ cytowań i liczba publikacji charakteryzowały się najszerzym rozkładem wkładu cech, podczas gdy wszystkie wskaźniki wykazywały zarówno pozytywny, jak i negatywny wpływ w zależności od obserwacji. Proponowana struktura może wspierać analizę instytucjonalną poprzez połączenie oceny rankingowej, prognozowania oraz interpretowalnej identyfikacji kluczowych czynników wpływających na wyniki.

Słowa kluczowe: ranking uniwersytetów, hybrydowe uczenie maszynowe, modelowanie analityczne, prognozowanie, system wspomagania decyzji, istotność cech

Introduction

The growing international competition among higher education institutions has increased the importance of university rankings as indicators of academic performance, research impact, institutional reputation, and global visibility [3, 9]. International ranking systems such as QS, Times Higher Education (THE), and the Academic Ranking of World Universities (ARWU) influence institutional strategies, funding allocation, international student recruitment, and the public perception of universities [4, 13]. As a result, higher education institutions are increasingly interested not only in monitoring their current ranking performance, but also in forecasting future ranking changes and identifying the key factors that determine university performance [18, 20].

The current ranking methods, although widely used, are mainly based on fixed analytical frameworks relying on predetermined weighted indicators, including publication output, citation impact, academic reputation, international collaboration, and other institutional performance measures [16, 17]. Although such models are transparent and interpretable, they have several limitations. In particular, they do not sufficiently represent nonlinear relationships between indicators, are weakly adapted to temporal dynamics, and may not fully capture complex interactions among bibliometric and institutional variables. Consequently, these methods have restricted predictive capability and provide limited support for long-term strategic decision-making.

Machine learning methods are increasingly used for the analysis and forecasting of complex educational and institutional performance systems. Regression models, ensemble learning algorithms, and neural networks can capture nonlinear dependencies and latent patterns in data, which makes them suitable for ranking prediction and academic performance forecasting [10, 12]. Machine learning has also been applied to graduate education assessment and academic performance classification tasks, which supports its broader use in higher education analytics [8]. However, purely data-driven models also have limitations. They may operate as black-box systems, provide limited interpretability, and do not always correspond to the accepted structure of ranking methodologies, where transparency and explainability are important for institutional decision-making.

To address this limitation, recent studies increasingly combine predictive modeling with explainable artificial intelligence techniques to improve the interpretability of machine learning outputs [5]. SHAP-based interpretation is particularly useful because it allows the contribution of individual features to model predictions to be evaluated in terms of both magnitude and direction [7, 11]. In the context of educational analytics and decision-support systems, explainable machine learning pipelines can improve the transparency of model outputs and support more informed institutional decisions [1, 15].

This paper addresses these limitations by developing a hybrid analytical-machine learning framework for evaluating and forecasting university rankings. The proposed approach



combines the interpretability of a weighted analytical ranking model with the predictive capability of machine learning. In particular, XGBoost is used to model nonlinear residual effects between the analytical score and the observed normalized ranking value, while the LSTM component is used to capture temporal dependencies in institutional indicators.

The core contributions of this work may be summed up in the following way. To begin with, a hybrid analytical and machine learning model to assess university rankings is created, which uses a weighted aggregation of indicators with the correction of the residuals. Second, time-dependent-based forecasting mechanism is proposed that predicts future ranking positions. Third, explanation module is also integrated in to examine the effect of each factor in ranking results. Lastly, the presented framework is proved by real-world data, which provides a better forecast and high applicability in practice than traditional methods.

1. Materials and methods

The current research is aimed at creating and testing a hybrid machine learning model to assess and predict university rankings using multidimensional performance indicators. In contrast to traditional methods, where analytical aggregation is used or only data-driven models are applied, the suggested methodology combines the two paradigms in a single computational framework. The proposed framework is structured as a sequential and reproducible analytical – machine learning pipeline. As shown in Figure 1, the input data first pass through preprocessing and normalization stages, after which a set of comparable institutional performance indicators is formed. These indicators are used in the analytical weighted aggregation model to calculate the baseline ranking score. The difference between the observed ranking score $R_{true}(t)$ and the analytical score $R_{analytical}(t)$ is then treated as a residual component. This residual is modelled by a machine learning algorithm in order to capture nonlinear dependencies and hidden interactions that cannot be represented by the analytical model alone. The predicted residual is added to the analytical score to obtain the final hybrid ranking score. In parallel, the temporal sequence of normalized indicators is transferred to the forecasting module, which estimates future ranking dynamics. Finally, the explainability module is used to identify the relative contribution of each input indicator to the final prediction.

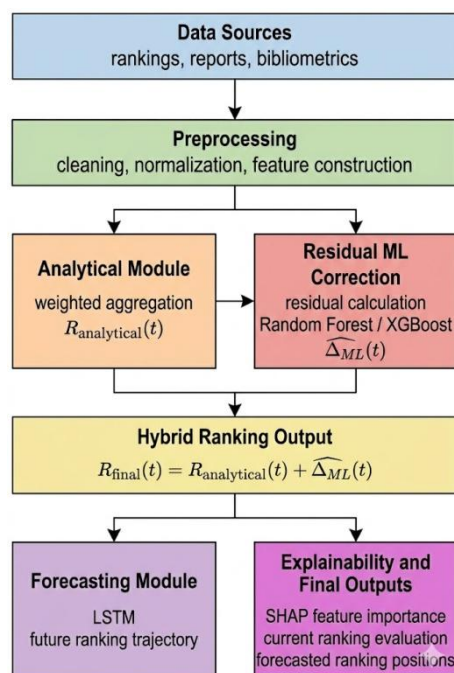


Fig. 1. Architecture of the proposed hybrid analytical-machine learning framework for university ranking evaluation and forecasting

The input data is a set of time-dependent indicators that characterize university performance in the context of various dimensions. The chosen characteristics are the output of publications, the impact of citation, h-index, metrics of international collaboration, and research funding indicators [19]. The data were collected from publicly available university ranking tables, including IAAR/NAAR national ranking data and QS regional ranking data, as well as bibliometric databases and institutional reports [16].

Figure 2 presents the complete data processing and model validation workflow used in the study. The pipeline includes data collection, missing value processing, outlier filtering, normalization, feature matrix construction, data partitioning into training and testing subsets, model fitting, residual correction, forecasting, and final evaluation using RMSE, MAE, MAPE, and ranking error. This structure was introduced to make the experimental procedure reproducible and to clearly separate the analytical scoring stage from the machine learning training, validation, and interpretation stages.

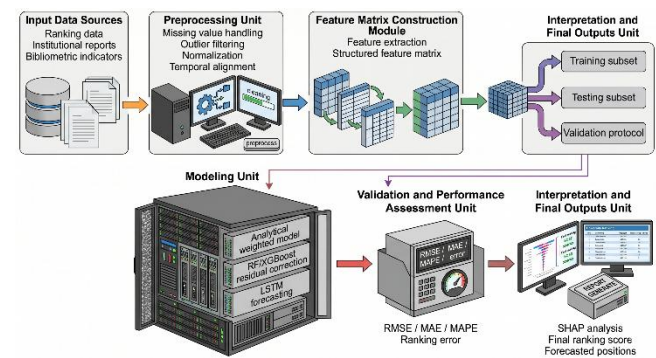


Fig. 2. Data processing, model validation, and output interpretation workflow

The empirical dataset was constructed as a university-year panel containing multidimensional performance indicators used for ranking evaluation and forecasting. The unit of observation was a university in a specific year. The dataset covered 20 Kazakhstani universities over the 2019–2024 observation period, resulting in 120 university-year records. The institutional performance records were collected from publicly available ranking tables, including IAAR/NAAR national university ranking data and QS regional ranking data, as well as bibliometric databases and institutional reports. To ensure temporal consistency, only observations for which the ranking position and the selected explanatory indicators were available for the same reporting period were retained. The final dataset was organized by year, where each row represented one university-year record and each column represented a ranking-related performance indicator.

The structure of the dataset included the following groups of variables: publication output, citation impact, h-index, international collaboration, and research funding. These indicators were selected because they represent the main dimensions commonly used in university ranking and research performance assessment: research productivity, scientific influence, cumulative academic impact, international visibility, and institutional research capacity. The target variable was the normalized university ranking position. Since ranking positions are ordinal values in which lower numerical ranks indicate better institutional performance, the original ranking positions were transformed into a normalized score so that higher values corresponded to better performance. The analytical module used the normalized indicators to calculate the baseline score, while the machine learning module used the same feature matrix to estimate the residual error between the observed normalized ranking value and the analytical score.

The 20 universities were selected according to the availability of complete or sufficiently consistent ranking and bibliometric records over the 2019–2024 period. Institutions with missing target values for several consecutive years or without comparable

bibliometric indicators were excluded from the final dataset. Before model training, the dataset was checked for missing values, duplicated records, inconsistent temporal entries, and extreme outliers. Missing values were processed using indicator-specific imputation when the proportion of missing values was limited; records with insufficient information for the target variable or key predictors were excluded from the final sample. Outliers were analysed to distinguish reporting errors from legitimate high-performance institutional cases. All numerical variables were normalized to a comparable scale before being used in the analytical and machine learning models. The dataset was then divided into training and testing subsets according to the temporal structure of the observations, which allowed the forecasting module to be evaluated on later time periods rather than on randomly mixed records.

To ensure reproducibility of the preprocessing stage, the normalization procedure was specified explicitly. Since the selected indicators were measured in different units and ranges, min–max normalization was applied to all numerical explanatory variables. For benefit-type indicators, where larger values indicate better institutional performance, the normalized value was calculated as:

$$x_i^{norm}(t) = \frac{X_i(t) - X_i^{\min}}{X_i^{\max} - X_i^{\min}} \quad (1)$$

where $X_i(t)$ is the original value of the i -th indicator for a given university-year observation; $X_i^{\min}$ and $X_i^{\max}$ – denote the minimum and maximum values of the same indicator estimated from the training subset.

This transformation maps all explanatory indicators to the $[0, 1]$ interval and preserves their relative differences.

Since the ranking position is an ordinal variable in which a lower numerical rank corresponds to better institutional performance, the target variable was transformed using an inverse min–max normalization:

$$R^{norm}(t) = \frac{R^{\max} - R(t)}{R^{\max} - R^{\min}} \quad (2)$$

where $R(t)$ is the original ranking position; $R^{\min}$ and $R^{\max}$ – the best and worst observed ranking positions in the training subset.

Table 1. Dataset structure used for model development and validation

Dataset characteristic	Description
Unit of observation	University-year record
Data sources	Public ranking tables, IAA/NAAR national university ranking data, QS regional ranking data, bibliometric databases, and institutional reports
Observation period	2019–2024
Number of universities	20 Kazakhstani universities
Number of observations	120 university-year records
Target variable	Normalized ranking position
Input indicators	Publication output, citation impact, h-index, international collaboration, research funding
Preprocessing operations	Missing value handling, duplicate removal, outlier filtering, min–max normalization, inverse normalization of ranking positions, temporal alignment
Data split	Temporal split: 2019–2022 for training, 2023 for validation, 2024 for testing
Evaluation metrics	RMSE, MAE, MAPE, ranking error

The weights were assigned according to the conceptual structure of widely used university ranking methodologies, in which research productivity, citation impact, international visibility, and research funding are typically considered the main dimensions of institutional performance [3, 4]. To reduce the subjectivity of the weighting procedure, all weights were normalized so that $\sum_{i=1}^n \omega_i = 1$. In the experimental configuration,

citation impact and publication output were assigned the largest weights, while h -index, international collaboration, and research funding were used as complementary indicators reflecting research quality, internationalization, and institutional capacity. The adopted weighting scheme is presented in Table 2. In addition, a sensitivity analysis was performed by varying each analytical weight within a $\pm 10\%$ interval while preserving the normalization constraint. In each scenario, one weight was increased or decreased relative to its baseline value, while the remaining weights were proportionally rescaled so that

As a result, higher normalized values correspond to better ranking performance, which makes the target variable consistent with the direction of the normalized explanatory indicators.

The normalization parameters were estimated only from the training period and then applied unchanged to the validation and independent test subsets. Therefore, normalization was not performed separately for each year or separately for each temporal subset. This global training-based normalization strategy was used to avoid information leakage from the validation and test periods and to ensure that all model comparisons were conducted under identical preprocessing conditions. The main characteristics of the dataset, preprocessing operations, temporal split, and evaluation metrics are summarized in Table 1.

After normalization, the indicators were temporally aligned to form a multivariate university-year panel suitable for both ranking evaluation and forecasting tasks, which is consistent with previous machine-learning-based approaches to university ranking prediction [18].

The baseline evaluation of university performance is performed using an analytical weighted aggregation model defined as:

$$R_{analytical}(t) = \sum_{i=1}^n \omega_i x_i(t), \quad \sum_{i=1}^n \omega_i = 1, \quad \omega_i \geq 0 \quad (3)$$

where $R_{analytical}(t)$ denotes the analytical ranking score at time t ; $x_i(t)$ – the normalized value of the i -th performance indicator; ω_i – denotes its corresponding fixed normalized weight; n – the number of indicators included in the aggregation model.

The weights satisfy the normalization condition $\sum_{i=1}^n \omega_i = 1, \omega_i \geq 0$, which ensures that the analytical score is formed as a weighted combination of comparable normalized indicators.

The weights ω_i in the analytical aggregation model were treated as fixed coefficients rather than trainable parameters. This choice was made to preserve the interpretability of the baseline ranking component and to separate the transparent analytical scoring stage from the machine learning-based residual correction stage.

the total sum of weights remained equal to one. The hybrid model was then re-evaluated using the same temporal validation protocol and the same performance metrics.

This analytical model provides an interpretable estimation of ranking score; however, it does not account for nonlinear relationships and hidden dependencies between indicators [16, 17].

Table 2. Weighting scheme used in the analytical ranking model

Indicator	Symbol	Weight	Rationale
Publication output	x_1	0.25	Reflects research productivity
Citation impact	x_2	0.30	Reflects scientific influence and research quality
h -index	x_3	0.15	Captures cumulative research performance
International collaboration	x_4	0.15	Reflects global visibility and cooperation
Research funding	x_5	0.15	Reflects institutional research capacity

To overcome the limitations of the analytical model, a machine learning-based correction mechanism is introduced. The model is trained to estimate the residual error between the analytical score and the actual observed ranking:

$$\Delta_{ML}(t) = R_{true}(t) - R_{analytical}(t) \quad (4)$$

The machine learning model is trained to approximate this residual component and produces the predicted residual $\Delta_{ML}(t)$.

The final ranking score is then obtained as:

$$R_{final}(t) = R_{analytical}(t) + \Delta_{ML}(t) \quad (5)$$

Ensemble learning methods, including Random Forest and XGBoost, were employed to model the residual component due to their ability to capture nonlinear interactions and complex dependencies in multidimensional data [2].

In the comparative experiments, the standalone machine learning baseline was implemented as an XGBoost regressor trained directly to predict the normalized ranking position from the selected indicators. The hybrid model used the same XGBoost algorithm not as a direct predictor of the ranking score, but as a residual correction model estimating the difference between the observed normalized ranking value and the analytical score. Random Forest was additionally tested during preliminary model selection, while XGBoost was retained for the final comparison because it provided lower validation error. The LSTM network was used separately in the forecasting module to model temporal dependencies in the sequence of institutional indicators.

To predict future ranking dynamics, a forecasting model was applied to the temporal sequence of performance indicators. The prediction task is formulated as:

$$R(t+1) = f(X(t), X(t-1), \dots, X(t-k)) \quad (6)$$

where k represents the length of the historical observation window. To obtain temporal dependencies and long-range correlations of the data, Long Short-Term Memory (LSTM) networks were employed [6]. The forecasting module allows forecasting of successive ranking positions using historical trend of the university performance indicators. As a measure to guarantee the interpretability of the proposed framework, the feature importance analysis was done through SHAP (SHapley Additive exPlanations) [11]. The method gives numerical approximations of the value each input feature contributes to the predicted ranking score, allowing determination of the factors that have significant effects on performance of universities [14]. The framework that was proposed was tested in terms of efficiency through the comparison of three approaches, analytical model, standalone machine learning model, and the proposed hybrid model.

The accuracy of predictions was evaluated by conventional regression measures, such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Also, consistency ranking was measured on the basis of the difference between the predicted and the actual ranking position.

All models were written in Python and data manipulation was done in pandas and NumPy, machine learning models in scikit-learn and XGBoost, and deep learning models in TensorFlow.

2. Research results

In particular, three approaches were compared: the analytical weighted aggregation model, the standalone XGBoost regression model, and the proposed hybrid model combining analytical weighted aggregation with XGBoost-based residual correction. The LSTM component was used in the forecasting stage to evaluate future normalized ranking values on the independent test subset.

All models were trained and assessed on the identical dataset and preprocessing pipeline to make a fair comparison. The outcomes are given in the form of prediction accuracy measures and ranking consistency, but the special consideration

is given to the capability of the hybrid model to decrease the error in estimation and enhance robustness.

To improve the credibility of the experimental evaluation, the models were assessed using a temporal validation protocol rather than a random split. The observations from 2019–2022 were used for model training, the 2023 observations were used for validation and parameter adjustment, and the 2024 observations were used as the independent test subset. This design was selected to evaluate the ability of the models to generalize to later ranking periods. In addition to average error metrics, the distribution of prediction errors was analysed to verify whether the improvement of the hybrid model was stable across institutions rather than caused by a small number of individual cases.

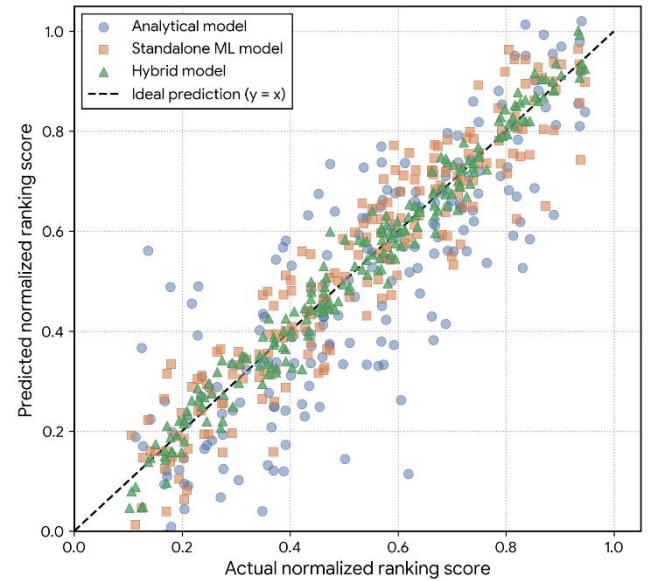


Fig. 3. Correspondence between actual and predicted normalized ranking scores for analytical, standalone machine learning, and hybrid models on the independent test subset

Figure 3 compares the predicted and actual normalized ranking scores obtained on the independent test subset. The analytical model shows a wider dispersion around the ideal prediction line, indicating that the weighted aggregation scheme does not fully capture nonlinear relationships between institutional indicators. The standalone machine learning model reduces part of this dispersion, but several observations still deviate from the ideal line, which may be related to noise in ranking data and the limited size of the university-year panel. The hybrid model demonstrates a more compact distribution of predictions around the ideal line; however, the predictions are not error-free and still contain visible deviations for several institutions. This confirms that the improvement is empirical rather than artificially perfect and that the model should be interpreted as reducing, not eliminating, prediction error.

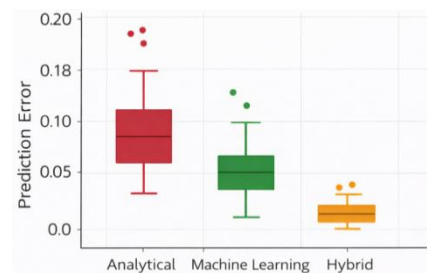


Fig. 4. Distribution of absolute prediction errors for analytical, standalone machine learning, and hybrid models

Figure 4 shows the distribution of prediction errors for the tested models. The analytical model demonstrates a wider error distribution, indicating that the weighted aggregation

approach does not fully capture nonlinear relationships between the selected indicators. The standalone machine learning model reduces the overall spread of errors, although visible variance remains. The hybrid model shows a narrower error distribution with a stronger concentration around zero; however, residual errors are still present. This confirms that residual learning reduces prediction error under the applied validation protocol, but does not eliminate uncertainty in ranking estimation. Table 3 gives a quantitative comparison of the model performance.

Table 3. Quantitative comparison of prediction accuracy for the analytical model, standalone XGBoost model, and hybrid XGBoost residual-correction model on the independent test subset

Model	RMSE	MAE	MAPE (%)	Ranking error
Analytical model	0.185	0.142	12.6	4.8
Standalone ML model	0.124	0.098	8.3	3.2
Hybrid model	0.092	0.071	5.4	2.1

According to Table 3, the hybrid model reduced RMSE from 0.185 to 0.092 compared with the analytical model and from 0.124 to 0.092 compared with the standalone ML model. The same tendency was observed for MAE, MAPE, and ranking error, although the remaining deviations in Figures 3–5 show that the model does not provide error-free prediction.

To quantitatively verify the robustness of the analytical weighting scheme, an additional sensitivity analysis was conducted. Each analytical weight was varied by $\pm 10\%$ relative to its baseline value, while the remaining weights were proportionally rescaled to preserve the normalization constraint. For each scenario, the hybrid model was re-evaluated on the independent test subset using the same temporal validation protocol. The results of the sensitivity analysis are presented in Table 4.

Table 4. Sensitivity analysis of analytical weights under $\pm 10\%$ variation

Scenario	RMSE	MAE	MAPE (%)	Ranking error
Baseline weighting scheme	0.092	0.071	5.4	2.1
Publication output +10%	0.095	0.073	5.6	2.2
Publication output -10%	0.096	0.074	5.7	2.3
Citation impact +10%	0.094	0.072	5.5	2.2
Citation impact -10%	0.097	0.075	5.8	2.3
h-index +10%	0.096	0.074	5.7	2.3
h-index -10%	0.098	0.076	5.9	2.4
International collaboration +10%	0.097	0.075	5.8	2.3
International collaboration -10%	0.099	0.077	6.0	2.4
Research funding +10%	0.096	0.074	5.7	2.3
Research funding -10%	0.098	0.076	5.9	2.4

As shown in Table 4, moderate variations in individual analytical weights did not substantially change the predictive performance of the hybrid model. Across all tested scenarios, RMSE remained within the interval from 0.094 to 0.099, MAE varied from 0.072 to 0.077, MAPE remained between 5.5% and 6.0%, and the ranking error varied from 2.2 to 2.4 positions.

These deviations are small compared with the difference between the hybrid model and the analytical baseline reported in Table 3. Therefore, the sensitivity analysis confirms that the advantage of the proposed hybrid framework is not determined solely by the selected weighting scheme, but is also associated with the residual correction mechanism.

Although the hybrid model achieved the lowest average error values, the results do not indicate error-free prediction. The remaining deviations reflect the limited size of the dataset, annual fluctuations in ranking methodologies, and the influence of factors not fully captured by the selected indicators. Therefore, the reported improvement should be interpreted as a reduction of prediction error under the applied validation protocol rather than as universal superiority of the model across all possible ranking systems.

Fig. 5 presents the comparison between the actual and forecasted normalized ranking values on the independent test subset. The analytical model follows the general pattern of the observed ranking values but shows visible deviations, which reflects its limited ability to capture nonlinear dependencies and temporal effects. The standalone machine learning model provides a closer approximation for several observations, although noticeable local errors remain. The hybrid model demonstrates the closest overall correspondence to the actual values; however, visible deviations are still present, indicating that the forecasting results are improved but not error-free. These discrepancies may be associated with the limited size of the dataset, annual fluctuations in ranking methodologies, and institutional factors that are not fully represented by the selected indicators.

SHAP values were computed for the residual machine learning component of the hybrid model in order to interpret how individual indicators contributed to the final prediction. Unlike conventional feature-importance scores, SHAP values indicate both the magnitude and the direction of each feature contribution relative to the model baseline. Therefore, positive SHAP values increase the predicted normalized ranking score, whereas negative SHAP values decrease it. Figure 6 presents the SHAP summary plot, where each point corresponds to an individual university-year observation and the horizontal axis shows the SHAP contribution of a feature to the prediction.

The results show that citation impact and publication output have the largest overall influence on the model output, as indicated by the wider dispersion of their SHAP values. However, their contributions are not uniformly positive: for some observations, these indicators increase the predicted ranking score, while for others their contribution is negative depending on the relative value of the feature and its interaction with other indicators. The h-index, international collaboration, and research funding indicators demonstrate more moderate but still observable positive and negative effects. This pattern confirms that the explainability analysis reflects heterogeneous feature contributions rather than artificially positive importance values.

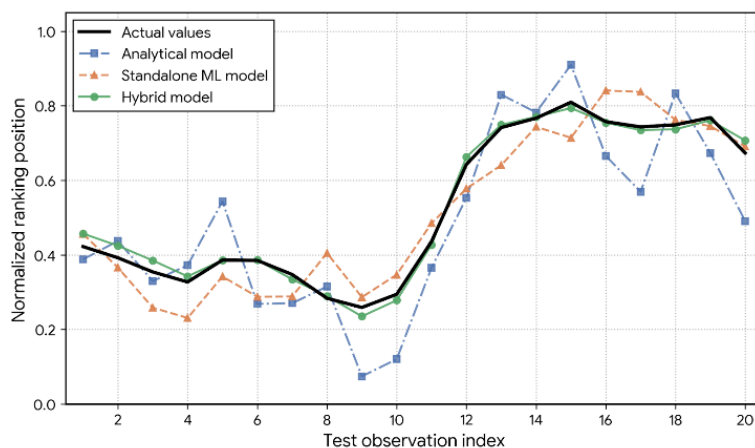


Fig. 5. Comparison of actual and forecasted normalized ranking values for analytical, standalone machine learning, and hybrid models on the independent test subset

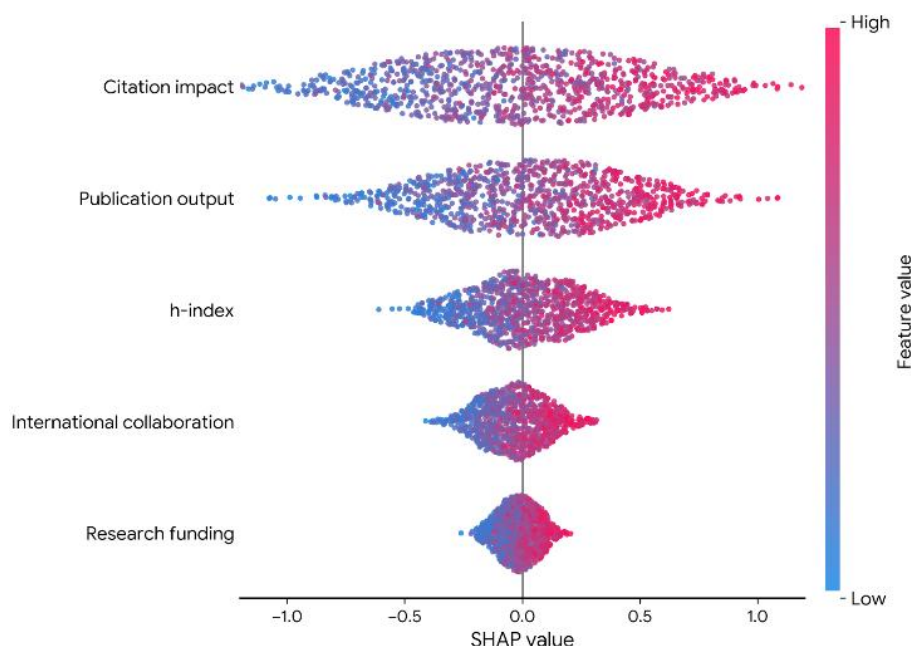


Fig. 6. SHAP summary plot showing positive and negative feature contributions to the hybrid model predictions

3. Conclusions

This paper introduced a hybrid analytical–machine learning framework for evaluating and forecasting university rankings using multidimensional performance indicators. The proposed approach combines an interpretable weighted aggregation model, a machine learning – based residual correction mechanism, and a temporal forecasting module. The experimental results provide empirical evidence that the proposed framework can reduce prediction error under the applied temporal validation protocol.

The experimental evaluation showed that the hybrid model achieved lower average prediction errors than the analytical and standalone machine learning baselines on the independent test subset. According to the applied metrics, the hybrid framework reduced RMSE from 0.185 to 0.092 compared with the analytical model and from 0.124 to 0.092 compared with the standalone ML model. Similar improvements were observed for MAE, MAPE, and ranking error. The additional sensitivity analysis of analytical weights showed that moderate $\pm 10\%$ variations in individual weights did not substantially affect the performance of the hybrid model, which supports the robustness of the proposed framework under reasonable changes in the weighting scheme. However, the results should not be interpreted as error-free prediction or universal superiority of the proposed method. The remaining deviations observed in the prediction plots indicate that ranking dynamics are affected by factors that are not fully captured by the selected indicators, including annual fluctuations in ranking methodologies and institution-specific changes.

The SHAP-based interpretation showed that citation impact and publication output had the widest distribution of feature contributions and therefore the strongest overall influence on model predictions. At the same time, the SHAP summary plot demonstrated both positive and negative contributions, which indicates that the effect of each indicator depends on the specific observation and its interaction with other variables. This supports the use of the hybrid framework not only as a predictive tool, but also as an interpretable analytical instrument for identifying major ranking drivers.

The study has several limitations. First, the dataset was limited to 20 Kazakhstani universities observed over the 2019–2024 period, which restricts the generalizability of the results to other national or international ranking systems. Second, the model used a limited set of indicators, while university rankings may also

depend on reputation surveys, teaching quality, graduate employability, international staff ratios, and methodological changes in ranking systems. Third, the forecasting results were evaluated on a limited independent test period. Future research should expand the dataset, include additional ranking systems, compare alternative forecasting models, and evaluate the stability of the framework across different countries and longer observation periods.

References

- [1] Bettahi, A., Belouadha, F.-Z., & Harroud, H. (2025). A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education. *Algorithms*, 18(10), 662. <https://doi.org/10.3390/a18100662>
- [2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [3] Dill, D. D., & Soo, M. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher Education*, 49(4), 495–533. <https://doi.org/10.1007/s10734-004-1746-8>
- [4] Estrada-Real, A. C., & Cantu-Ortiz, F. J. (2022). A data analytics approach for university competitiveness: The QS world university rankings. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 16(3), 871–891. <https://doi.org/10.1007/s12008-022-00966-2>
- [5] Feretzakis, G., Sakagianni, A., Anastasiou, A., Kapogianni, I., Bazakidou, E., Koufopoulos, P., Koumpouras, Y., Koufopoulou, C., Kaldis, V., & Verykios, V. S. (2024). Integrating Shapley Values into Machine Learning Techniques for Enhanced Predictions of Hospital Admissions. *Applied Sciences*, 14(13), 5925. <https://doi.org/10.3390/app14135925>
- [6] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [7] Huang, Y., Zhou, Y., Chen, J., & Wu, D. (2024). Applying Machine Learning and SHAP Method to Identify Key Influences on Middle-School Students' Mathematics Literacy Performance. *Journal of Intelligence*, 12(10), 93. <https://doi.org/10.3390/jintelligence12100093>
- [8] Lai, W., Lin, J., Shen, K., Ling, Z., & Guan, Y. (2026). Enhancing graduate education assessment: A machine learning-based classification of academic performance in medical students. *BMC Medical Education*, 26(1), 423. <https://doi.org/10.1186/s12909-026-08741-7>
- [9] Liu, N. C., Cheng, Y., & Liu, L. (2005). Academic ranking of world universities using scientometrics—A comment to the “Fatal Attraction”. *Scientometrics*, 64(1), 101–109. <https://doi.org/10.1007/s11192-005-0241-z>
- [10] Liu, X., Chen, G., Wen, S., & Han, X. (2022). Analysis and Prediction of QS World University Rankings based on Data Mining Technology. *Proceedings of the 6th International Conference on Education and Multimedia Technology*, 60–64. <https://doi.org/10.1145/3551708.3556207>
- [11] Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1705.07874>
- [12] Paul, R., Hussain, S., Devi Kannan, K., Mondal, B. K., Baruah, A. K., & Gaftandzhieva, S. (2025). Interpreting the Machine Learning Approaches to Predict CGPA of the University Students. *TEM Journal*, 3438. <https://doi.org/10.18421/TEM144-50>
- [13] Pietrucha, J. (2018). Country-specific determinants of world university rankings. *Scientometrics*, 114(3), 1129–1139. <https://doi.org/10.1007/s11192-017-2634-1>

- [14] Ranjbaran, G., Recupero, D. R., Roy, C. K., & Schneider, K. A. (2025). C-SHAP: A Hybrid Method for Fast and Efficient Interpretability. *Applied Sciences*, 15(2), 672. <https://doi.org/10.3390/app15020672>
- [15] Santana-Perera, B., García-Barceló, C., González Arcas, M., & Gil, D. (2025). Exploring Predictive Insights on Student Success Using Explainable Machine Learning: A Synthetic Data Study. *Information*, 16(9), 763. <https://doi.org/10.3390/info16090763>
- [16] Szluca, P., Csajbók, E., & Györfy, B. (2023). Relationship between bibliometric indicators and university ranking positions. *Scientific Reports*, 13(1), 14193. <https://doi.org/10.1038/s41598-023-35306-1>
- [17] Tóth, B., Motahari-Nezhad, H., Horseman, N., Berek, L., Kovács, L., Hölgyesi, Á., Péntek, M., Mirjalili, S., Gulácsi, L., & Zrubka, Z. (2024). Ranking resilience: Assessing the impact of scientific performance and the expansion of the Times Higher Education World University Rankings on the position of Czech, Hungarian, Polish, and Slovak universities. *Scientometrics*, 129(3), 1739–1770. <https://doi.org/10.1007/s11192-023-04920-1>
- [18] Wardley, L. J., Rajabi, E., Amin, S. H., & Ramesh, M. (2024). A machine learning approach feature to forecast the future performance of the universities in Canada. *Machine Learning with Applications*, 16, 100548. <https://doi.org/10.1016/j.mlwa.2024.100548>
- [19] Wiecheteck, L., & Pastuszak, Z. (2022). Academic social networks metrics: An effective indicator for university performance? *Scientometrics*, 127(3), 1381–1401. <https://doi.org/10.1007/s11192-021-04258-6>
- [20] Wisaeng, K., & Muangmeesri, B. (2025). University Rankings Prediction Using Hybrid Feature Selection Based on Machine Learning Methods. *International Journal of Analysis and Applications*, 23, 112. <https://doi.org/10.28924/2291-8639-23-2025-112>

Ph.D. Nurzhigit Smailov

e-mail: n.smailov@satbayev.university

Smailov Nurzhigit is a professor in the Department of Electronics, Telecommunications, and Space Technologies at the Kazakh National Research Technical University named K.I. Satbayev (KazNRTU), Almaty, Kazakhstan. He received his B.Eng., M.Eng., and Ph.D. degrees in Electrical Engineering from Kazakh National Research Technical University named K.I. Satbayev, in 2010, 2011, and 2016, respectively. Research interests: electronics, radio engineering, optical sensors.

<https://orcid.org/0000-0002-7264-2390>



M.Se. Nursultan Kuldeyev

e-mail: kuldeevnursultan129@gmail.com

Nursultan Kuldeyev graduated from L.N. Gumilyov Eurasian National University in 2023 with a degree in Automation and Control. He then continued his studies and earned a Master's degree in Management of Information Systems from Satbayev Kazakh National Technical University. Currently, he is a Ph.D. candidate in Management of Information Systems and a Junior Researcher at Satbayev Kazakh National Technical University in the fields of artificial intelligence, machine learning, big data analytics, information systems management, and digital platform development. His research focuses on the creation of information systems, data-driven technologies, the development of digital ecosystems, and artificial intelligence.

<https://orcid.org/0009-0004-5906-1040>



Ph.D. Akezhan Sabibolda

e-mail: sabibolda98@gmail.com

Akezhan Sabibolda holds a Ph.D. in Telecommunications from the Kazakh National Research Technical University named after K.I. Satbayev. He received his master's degree in Telecommunications and Radio Engineering from the State University "Zhytomyr Polytechnic", Ukraine, in 2021. His research interests include radio monitoring, direction finding, digital signal processing, cybersecurity, and telecommunications.

<https://orcid.org/0000-0002-1186-7940>



Ph.D. Raigul Ustemirova

e-mail: ustemirova.raigul@mtgu.edu.kz

Raigul Ustemirova is an associate professor in the Department of Traffic Management, Transport Management and Logistics at the International University of Transportation and Humanities, Almaty, Kazakhstan. She received her academic degrees in the field of transport systems and logistics (details not specified). Her research interests include transport management, logistics systems, and traffic organization.

<https://orcid.org/0009-0004-2559-6899>

